TECHNICAL DOCUMENT 3300
September 2015

Multistage Analysis of Cyber Threats for Quick Mission Impact Assessment (CyberIA)

Henry Au

SSC Pacific
San Diego, CA 92152-5001

SSC Pacific
San Diego, California 92152-5001

K. J. Rothenhaus, CAPT, USN
Commanding Officer

C. A. Keeney
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed by the ISR Engineering Branch (Code H56D0) of the C2 and Networks Division (Code H5300), SPAWAR Systems Activity Pacific (SSA Pacific), Honolulu, Hawaii. The Naval Innovative Science and Engineering (NISE) Program at Space and Naval Warfare Systems Center Pacific (SSC Pacific) funded this Applied Research project.

Released by
J. Lee, Head
ISR Engineering Branch

Under authority of
W. Fukumae, Head
C2 and Networks Division

ACKNOWLEDGEMENTS

This CyberIA project was funded under the internal Applied Research NISE Fiscal Year 2015 program managed by Robin Laird (Code 72120) and Dave Rees (Code 72120). This work was possible with the support of the NISE Program and its umbrella project management plan (PMP) document. The cyber community at SSC Pacific facilitated both growth and direction of this project. Team members Mamadou Diallo (Code 58240) and Krislin Lee (Code H56F0) provided their expertise in Web development and programming support for this project.

CONTENTS

Figures

Tables

1. INTRODUCTION

Current solutions rely on a combination of intrusion detection systems (IDS), an intrusion prevention system (IPS), and security information and event management (SIEM) technologies to identify cyber threats to network systems based on a host of physical and virtual network sensors. These traditional IDS/IPS and SIEM cyber security solutions often generate large sets of log data that can hide detected threats. As networks grow and the magnitude of generated alarms increases, analysts are faced with both big data storage and access problems. Access times become an issue. Analysts cannot prioritize threats and examine reoccurring threats. Figure 1 shows an IDS alarm generated by Snort®, an open-source IDS used for network alarm generation.

Algorithms, such as k-means clustering and support vector machines (SVM), can reduce the number of threats to a manageable level. With a manageable list of events, network value assets using a graph database, and an SVM to monitor behaviors long term, one can produce a system that reduces information overload. This system will leverage numerous established technologies and provide a better situational awareness of the monitored cyber system.

```
(Event)
        sensor id: 0    event id: 1     event second: 920898013 event microsecond: 194424
        sig id: 8       gen id: 125     revision: 1      classification: 3
        priority: 2     ip source: 206.48.44.18 ip destination: 172.16.112.100
        src port: 1054  dest port: 21   protocol: 6     impact_flag: 0  blocked: 0

Packet
        sensor id: 0    event id: 1     event second: 920898013
        packet second: 920898013        packet microsecond: 194424
        linktype: 1     packet_length: 80
[    0] 00 10 7B 38 46 32 00 C0 4F A3 58 23 08 00 45 00  ..{8F2..O.X#..E.
[   16] 00 42 9D 00 40 00 7F 06 47 FE CE 30 2C 12 AC 10  .B..@...G..0,...
[   32] 70 64 04 1E 00 15 00 17 AD 57 00 17 AF 17 50 18  pd.......W....P.
[   48] 21 A2 21 89 00 00 70 6F 72 74 20 31 39 39 2C 31  !.!...port 199,1
[   64] 39 39 2C 31 39 39 2C 31 39 39 2C 30 2C 38 30 0A  99,199,199,0,80.
```

Figure 1. Snort® alarm.

2. SYSTEM ARCHITECTURE

The CyberIA system architecture is built on a conventional client server model. The client can access the service through a Web page. The server handles the user request. In this case, the server handles the start and stop time of the log minimization system. Figure 2 shows the client–server architecture.

For CyberIA, an open-source C++ Web development framework was selected. This framework was selected due to its capability of handling high loads. This competency is achieved by using a modern C++ as the development language designed to develop both websites and Web services. The framework allows seamless integration of C++ libraries, thus providing the ideal framework for developing and integrating different high-performance C/C++ algorithms. This capability is significant because NVIDIA® CUDA™ architecture utilizes C/C++ coding to exploit the processing power of graphics processing units (GPUs) to port highly parallelizable and computationally expensive code to the graphical processing unit (GPU) for processing.

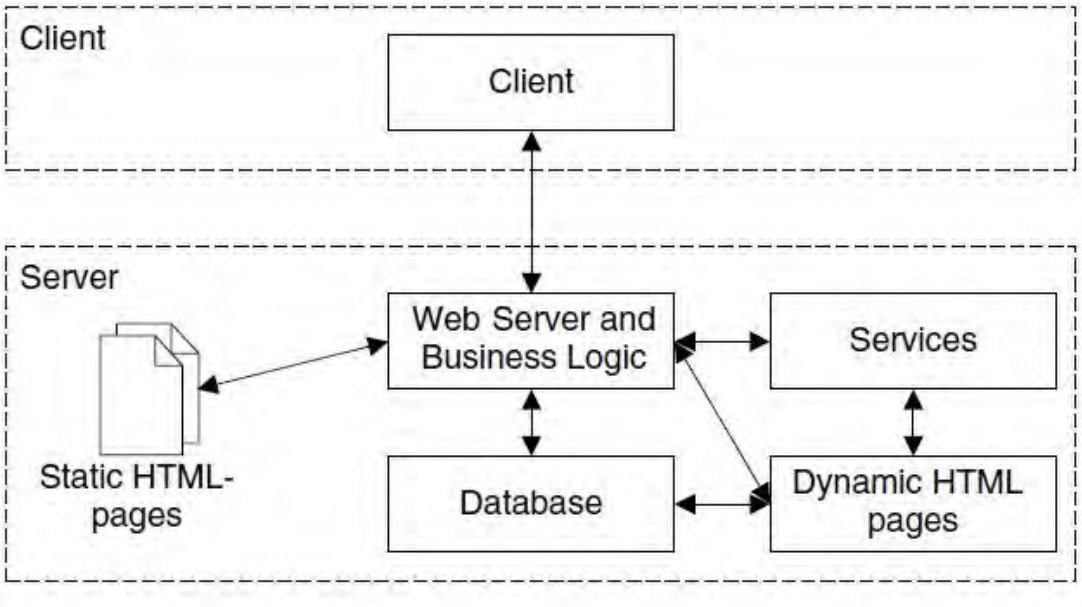

Figure 2. Conventional client–server Web architecture.

3. DATABASE AND SERVICES

To generate test data, the project team used the Defense Advanced Research Projects Agency's (DARPA) 1999 network intrusion data set (freely available) and labeled attacks. Snort® processed the DARPA network packet capture (pcap) data. Using Barnyard2, an open-source interpreter for Snort® unified2 binary output files, the binary data parsing and storage to disk is separated to another process that will not allow Snort® to miss network traffic. Alarms are saved into the commonly used open-source MySQL® database. Figure 3 describes the data flow process.

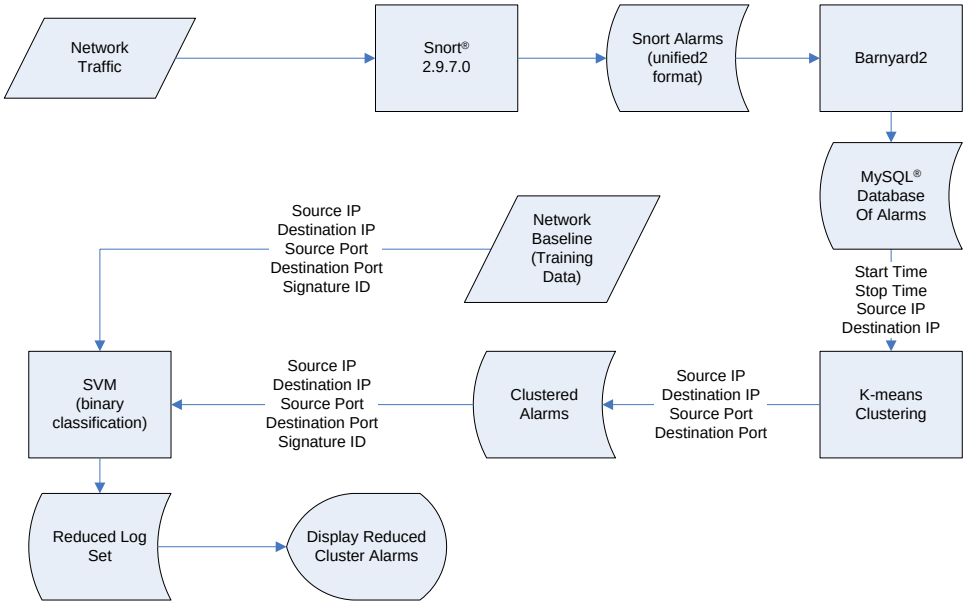

Figure 3. CyberIA data flow process.

The generated database is then processed using the first-phase k-means clustering. Results of clusters are further processed by a supervised machine-learning algorithm, SVM, which will binary classify alarms to minimize the false positive alarms. Results are then displayed to the user. Figures 4 and 5 also show (from a top and detailed level) the Snort® database schema used for data access.

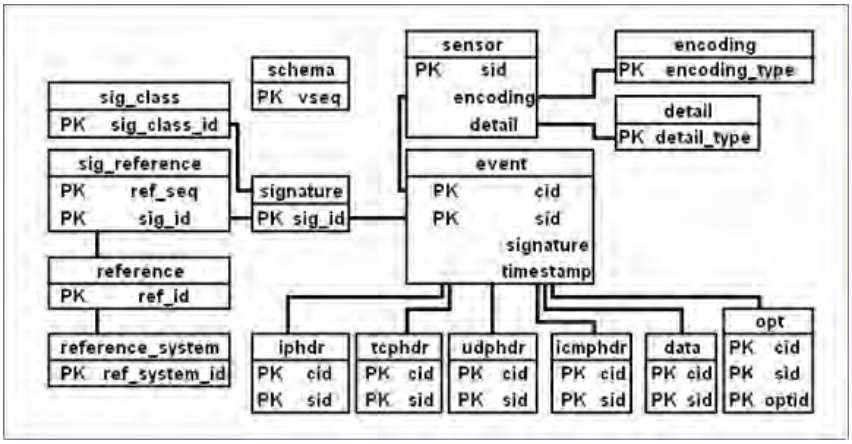

Figure 4. Top-level Snort® database schema.

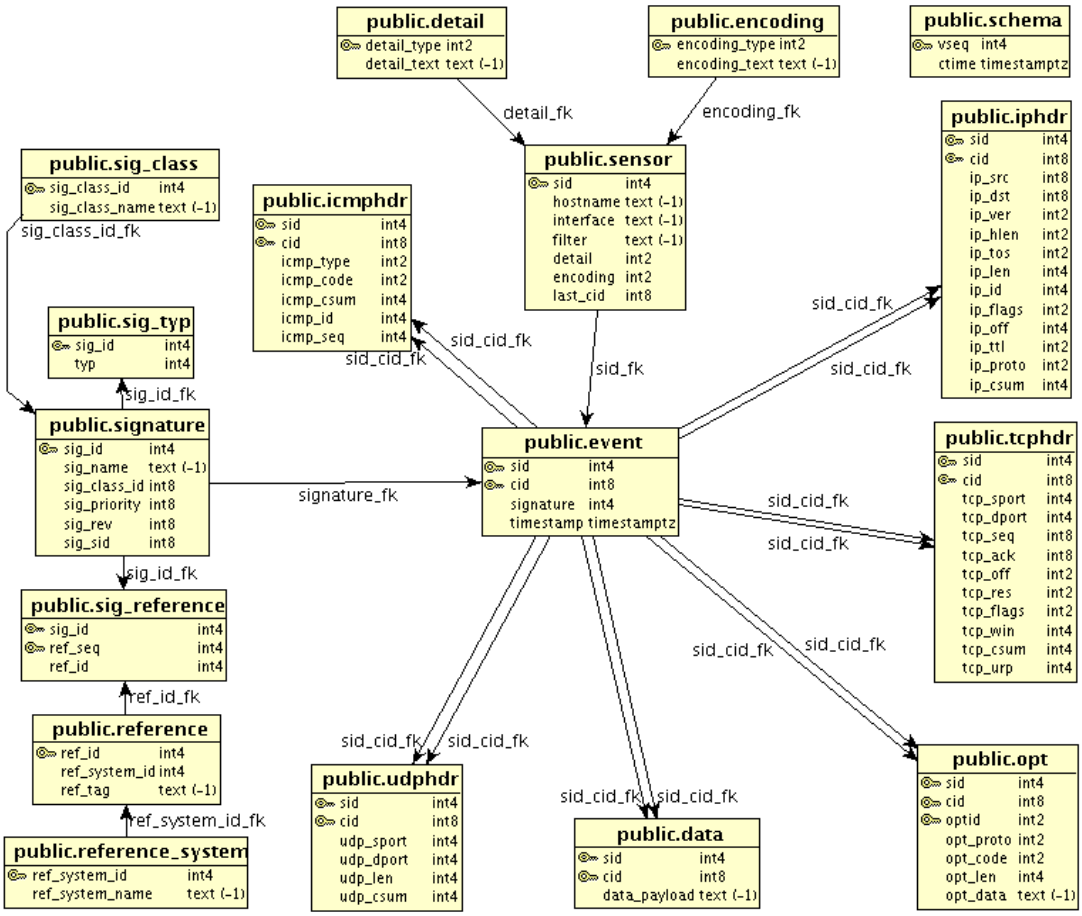

Figure 5. Detailed Snort® database schema.

3.1 PHASE-ONE ALGORITHM

Initially, CyberIA focused on the use of an unsupervised machine-learning algorithm to support the clustering of data, specifically self-organizing maps (SOMs); however, tests revealed that normalizing data to support a faster calculation produces a poorly clustered SOM. This method also made centroid determination more difficult. We were mapping from a three-dimensional (3-D) space to a visual two-feature representation; the three features were time, source (Internet Protocol) IP, and destination IP. With low-space mapping, clusters were identified with traditional image processing techniques. Normalizing data led to a loss of fidelity required for cyber forensics and the network graph database. As a result, CyberIA moved away from using a SOM for the first phase to a k-means algorithm. The k-means offered a better performance when an adequate value of k clusters than previously selected.

The parameters for the k-means clustering are as follows:

$$k = \text{numAlarms}/(\text{time window}*25) \quad (1)$$
$$\text{theta} = 2*\text{numAlarms}$$
$$\text{threshold} = 0.0002$$

3.2 PHASE-TWO ALGORITHM

An open-source C++ support vector machine implementation is set for use with binary classification to reduce the number of alarms presented to the user. We are currently adjusting the application to work with both the Snort® IDS alarm data and schema. The project team will also investigate and develop a good process for network baseline needs. This effort will continue in Fiscal Year (FY) 2016.

4. IMPLEMENTATION AND EVALUATION

In FY15, the project team produced a complete framework and an integrated first-phase clustering algorithm. Table 1 provides the k-means processing time for various 24-hour attack windows from the Snort® processed DARPA network pcap. Figure 6 shows the (almost) linear increase associated with the number of items and the clustering algorithm. The rise was expected as the calculation for k value was adjusted based on the number of alarms, and with an increase in k value, the processing time increased.

Table 1. K-means processing time using 24-hour Snort® alarm window.

Test #	# IDS Alarms	K-means Processing Time (ms)
1	8116	837
2	11952	1,815
3	12146	1,431
4	12777	2,626
5	13609	2,332
6	14331	2,598
7	21062	5,617
8	21724	6,027
9	22444	4,827
10	27076	6,978
11	28515	11,076
12	29914	8,625
13	31293	12,539
14	35765	12,138
15	51547	25,536

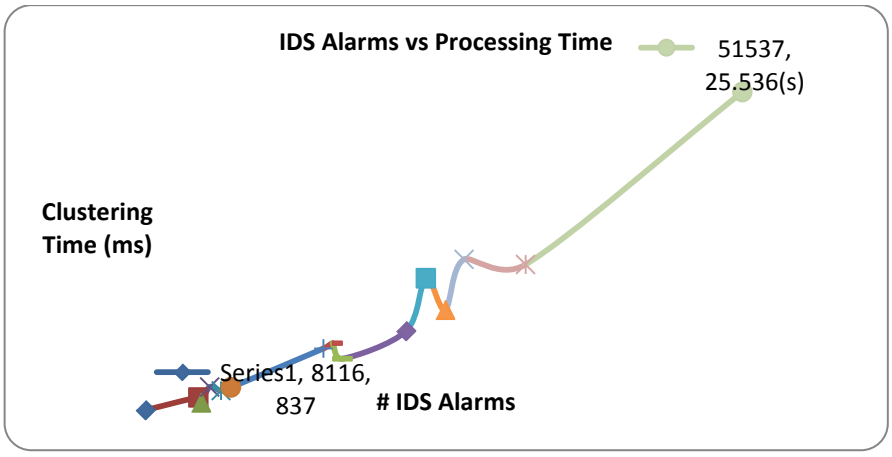

Figure 6. The number of IDS alarms vs. the k-means clustering processing time.

Timestamp, source IP, and destination IP (three selected features) have distinct meanings when using randomly generated centroids. Based on the k equation provided, we can further reduce the number of k clusters and apply them more effectively to cluster attack scenarios, which would reduce the process time. Figure 7 shows the implemented Web front end and the end-to-end system framework.

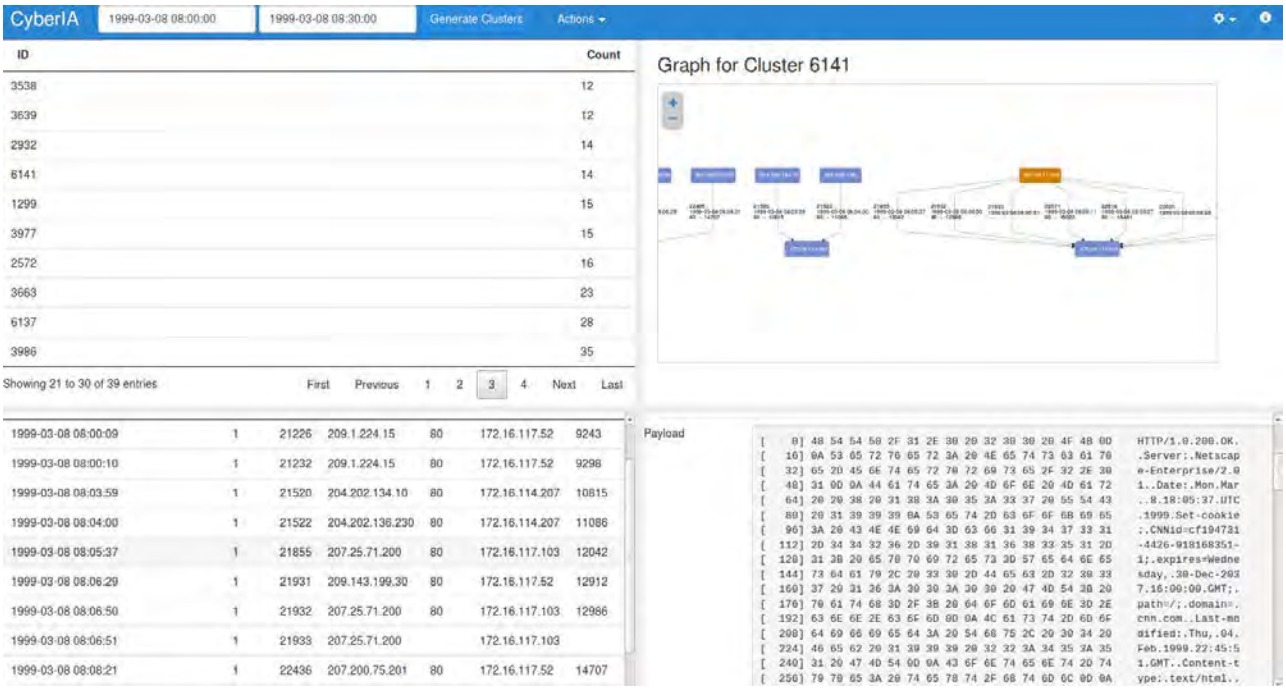

Figure 7. Completed CyberIA system Web access graphical user interface (GUI).

5. CONCLUSION

In this technical document we presented the proof-of-concept (POC) CyberIA system, a data-driven intrusion detection log analysis tool capable of processing thousands of logs. CyberIA makes use of a k-means clustering algorithm developed in house. The algorithm has integrated database access and a complete Web framework capable of integrating other C/C++ algorithms developed in house. The system allows for the ease of GPU integration to reduce the processing time, thus allowing the process of large data sets in real time.

6. FUTURE WORK

In FY16, CyberIA development will continue. The project team will focus on the centroid initialization for k-means clustering. Since IPs are distinct, using a randomly generated centroid may not be optimal during clustering. Without randomly selecting centroids, k-means clustering becomes a deterministic system. This will benefit users and simplify forensic analysis. The user is presented with consistent clusters when using the same parameters.

With the completion of the proof-of-concept framework, we can integrate the second phase supervised machine-learning algorithm and tuning using the recent (and available) network data from the University of New Brunswick Information Security Centre of Excellence (ISCX). Network data are labeled and contain more recent and complex cyber exploitations. The k-means algorithm for big data scalability is set for detailed timing analysis. We can port many algorithms (developed in house) onto the GPU to decrease processing time using the detailed timing analysis. The Davies–Bouldin Index (DBI) can help assess k-means clustering. To facilitate the mission impact assessment portion of the CyberIA framework, we will use the integration of a global network graph database for alarms.

BIBLIOGRAPHY

Abouabdalla, O., H. El-Taj, A. Manasrah, and S. Ramadass. 2009. “False Positive Reduction in Intrusion Detection System: A Survey.” *2nd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BMNT ’09)* (pp. 463–466). October 18–20, Beijing, China. IEEE.

Hachmi, F. and L. Mohamed. 2013. “A Two-Stage Technique to Improve Intrusion Detection Systems Based on Data Mining Algorithms.” *5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)* (pp. 1–6). April 28–30, Hammamet, Tunisia. IEEE.

Remya, R. and S. Anil. 2013. “A Hybrid Method Based on Genetic Algorithm, Self-Organised Feature Map, and Support Vector Machine for Better Network Anomaly Detection.” *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1–5). July 4–6. Tiruchengode, India. IEEE.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
September 2015	Final	

4. TITLE AND SUBTITLE

Multistage Analysis of Cyber Threats for Quick Mission Impact Assessment (CyberIA)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Henry Au

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SSC Pacific, 53560 Hull Street, San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

TD 3300

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Innovative Science and Engineering (NISE) Program (Applied Research)
SSC Pacific, 53560 Hull Street, San Diego, CA 92152–5001

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Network intrusion detection systems (IDS) are powerful network defense tools that monitor network traffic in real time and generate alarms based on known signatures; however, the increasing complexity of cyber threats (e.g., advanced malware), distributed denial-of-service attacks, and session-hijacking have produced large alarm sets. Analysts may miss an alarm or a mission-critical system may become compromised due to the amount of data required for processing. This information overload often leads to unknown cyber postures, system capabilities, and ultimately mission impacts due to cyber threats.

In this technical document, we propose Multistage Analysis of Cyber Threats for Quick Mission Impact Assessment (CyberIA), a multistage approach to log reduction as well as the development of framework to support IDS alarm analysis for network impact assessments. The system is composed of two phases of algorithms. The first phase utilizes a k-means clustering algorithm, and the second phase utilizes a supervised machine-learning system to minimize the clustered log sets. The final result is coupled with a network graph database to determine the impact on networked systems.

15. SUBJECT TERMS

Mission Area: Intrusion Detection Systems (IDS)

cyber threat	machine-learning algorithm	k-means clustering
support vector machines	self-organizing maps (SOMs)	big data

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Henry Au
U	U	U	U	16	**19B. TELEPHONE NUMBER** *(Include area code)* (808) 474-4179

INITIAL DISTRIBUTION

84300	Library	(2)
85300	Archive/Stock	(1)
H56D0	H. Au	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (1)

Approved for public release.

SSC Pacific
San Diego, CA 92152-5001